Published in final edited form as:

Dev Cell. 2015 March 23; 32(6): 772–774. doi:10.1016/j.devcel.2015.03.004.

Reassessment of Piwi binding to the genome and Piwi impact on RNA polymerase II distribution

Haifan Lin[1,2,*], **Mengjie Chen**[3], **Anshul Kundaje**[4,5], **Anton Valouev**[6], **Hang Yin**[7], **Na Liu**[1,2], **Nils Neuenkirchen**[1,2], **Mei Zhong**[1,2], and **Michael Snyder**[4]

[1]Yale Stem Cell Center, Yale University, New Haven, CT 06520, USA

[2]Department of Cell Biology, Yale University, New Haven, CT 06520, USA

[3]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, USA

[4]Department of Genetics, Stanford University, Stanford, California 94305, USA

[5]Department of Computer Science, Stanford University, Stanford, California 94305, USA

[6]Division of Bioinformatics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA

[7]Department of Department of Biochemistry & Molecular Biology, University of Georgia, Athens, GA30602, USA

Abstract

Drosophila Piwi was reported by Huang *et al*. (2013) to be guided by piRNAs to piRNA-complementary sites in the genome, which then recruits Heterochromatin Protein 1a and histone methyltransferase Su(Var)3-9 to the sites. Among additional findings, Huang *et al*. (2013) also reported Piwi binding sites in the genome and the reduction of RNA polymerase II in euchromatin but its increase in pericentric regions in *piwi* mutants. Marinov *et al*. (2015) disputed the validity of the Huang *et al*. bioinformatic pipeline that led to the last two claims. Here we report our independent reanalysis of the data using current bioinformatic methods. Our reanalysis agrees with Marinov *et al*. (2015) that Piwi's genomic targets still remain to be identified, yet confirms the Huang *et al*. claim that Piwi influences RNA polymerase II distribution in the genome. This Response addresses the Marinov *et al*. (2015) Matters Arising, published concurrently in *Developmental Cell*.

How epigenetic factors are recruited to specific sites in the genome represents a key question in epigenetic programming. Huang *et al*. (2013) reported the function of *Drosophila* Piwi, a member of the Argonaute protein family, and its cognate piRNAs in

---

*Correspondence: haifan.lin@yale.edu.

epigenetic factor recruitment. There are seven main claims by Huang *et al*. (2013): (1) piRNA is both necessary and sufficient to recruit Piwi to a piRNA-complementary site in the genome; (2) Heterochromatin Protein 1a (HP1a), a Piwi-interactor, and histone methyltransferase Su(Var)3-9 are also recruited to the piRNA complementary site, inducing its suppressive chromatin state; (3) such recruitment in euchromatin appears to be mediated by piRNA binding to nascent RNAs tethered to chromatin; (4) Piwi-piRNA binding to targets is highly sequence-specific, with one, two, and three mismatches reducing binding by ~40%, 60%, and 90%, respectively; (5) Piwi-piRNA complexes bind to numerous piRNA-complementary sites in the genome; (6) Piwi mutations influence the global distribution of epigenetic marks and reduce RNA polymerase II (Pol II) in euchromatin but increase it in pericentric regions enriched with transposons and repeats; and (7) Piwi functions in global transcriptional silencing of transposons.

Recently, Marinov *et al*. (2015) reanalyzed the ChIP-seq data by Huang *et al*. (2013) and disputed the validity of the bioinformatic pipeline used by Huang *et al*. (2013) in reaching Claim 5 and the Pol II part of Claim 6. We appreciate the effort of Marinov *et al*., and have independently reanalyzed these data using current bioinformatic tools. Our analysis is reported below.

Our reanalysis of the Piwi ChIP-seq data by Huang *et al*. (2013)

Claim 5 by Huang *et al*. (2013) was based on Piwi ChIP-seq analysis using an in-house pipeline (herein called “the Yin pipeline”) that did not involve peak calling, but only mapped sequence reads onto the genome, using enrichment scores to identify sequences that are likely bound by Piwi (Yin et al., 2011). In addition, this pipeline did not apportion for reads that map to repetitive sequences. Using this pipeline, Huang *et al*. (2013) concluded that Piwi is enriched in many piRNA-corresponding sites in the genome. When Marinov *et al*. used this pipeline to analyze the Piwi ChIP-seq data from Huang *et al*. (2013), it produced Piwi profiles essentially identical to Huang *et al*. (2013; *e.g.*, Figure 1B in Marinov *et al*., 2015). This indicates that no human error was introduced by Huang *et al*. in running this pipeline.

However, when Marinov *et al*. (2015) reanalyzed the same data using their pipeline (herein called “the Marinov pipeline”), they found no significant enrichment of Piwi at specific sites in the genome. To identify the cause of this striking difference, we compared the two pipelines. They had very different purposes: the Yin pipeline tried to capture signals from any possible repetitive region in order not to miss repetitive binding, while the Marinov pipeline weighted the ambiguous counting of repetitive sequences. Thus the Yin pipeline overcounted the repetitive region signals, while the Marinov pipeline undercounted the repetitive region signals. Piwi is very different from other chromatin binding factors such as transcription factors or modified histones (see below), for which appropriate programs have been specifically developed for detecting binding sites according to their binding features in the genome. We therefore used a biologically more meaningful algorithm, CSEM (ChIP-Seq multi-read allocation using Expectation-Maximization), that allocates multimapping reads as fractional counts according to the abundance of the unique reads at/near each of the repeat

sites (Chung et al., 2011). However, the CSEM algorithm did not reproduce the results of Huang *et al*. (2013).

We realized that at least part of the problem that prevented a more confident outcome in our reanalysis and that of Marinov *et al*. (2015) is the relatively low quality of the Piwi ChIP-seq data by Huang *et al*. (2013) as measured by today's standard. The data were obtained in 2007 using a first generation Solexa sequencer that generated a relatively small number of short reads with a high error rate (only the first 25 bp sequences are reliable), which necessitated more forgiving recursive alignment criteria (Huang *et al*., 2013). In order to obtain better sequence data from the original ChIP experiments, we recovered the minute amounts of the original samples by Huang *et al*. (2013) and re-sequenced them using HiSeq2000. Unfortunately, this did not improve the mappable rate, possibly due to deterioration during long-term storage and needed over-amplification. Therefore, we agree with Marinov *et al*. that Piwi's genomic targets still remain to be identified.

In our hands, Piwi-ChIP has been more difficult to achieve than ChIP of conventional epigenetic factors, marks, and RNA Pol II, and we are cognizant of the fact that it has not been achieved outside the Lin lab. Nevertheless, four out of six Lin lab members have achieved Piwi- ChIP with repeated success in different tissues and using different antibodies. It remains unclear why Piwi does not produce a robust ChIP signal. This might reflect certain peculiar aspects of Piwi-piRNA interaction with chromatin that are not suited for detection by the ChIP approach. Therefore, we have conducted DamID mapping of Piwi (to be published) and are continuing to test all relevant claims by Huang *et al*. (2013).

Our re-analysis of the Pol II ChIP-seq data by Huang *et al*. (2013)

Marinov *et al*. (2015) also disputed the claim by Huang *et al*. (2013) that in *piwi* mutants Pol II "is reduced in euchromatic regions but increased in pericentrosomal regions enriched with transposons and repeats". As discussed above, the Yin pipeline overcounted multimapping reads yet the Marinov pipeline used transposon consensus, instead of real transposon insertions, to map Pol II, which may miss discoveries. We therefore performed three independent analyses of the Pol II ChIP-seq data by Huang *et al*. (2013) by using transposon annotations on the fly genome. In the first two analyses, all reads were trimmed to 32 bp, and were mapped against the *Drosophila* genome (Flybase r5.22) allowing up to 2 mismatches but no indel. In the first analysis, we summarized the mapping information at the repeat name level. For each repeat name, RPKM and RPM scores were calculated for the quantification of abundance. Only reads unique to a specific repeat name were used in the calculation. If a read can be mapped to multiple loci and these loci correspond to non-overlapping repeat and genic regions, the read was not used in the analysis. As shown in Figures 1A-D, Pol II is enriched in many transposons and repeat sequences.

In the second analysis, we allocated reads using the CSEM algorithm (Chung et al., 2011), and performed peak calling using MOSAiCS (Sun et al., 2013) (Figure 1E). Finally we annotated all the peaks using FlyBase annotation. Overall, the binding of Pol II in the transposon and repeat sequences increased from 5% in the wildtype genome to 26% in the

piwi mutant genome, indicating potential enrichment of Pol II in these repeat and transposon sequences.

In the third analysis, we used read mapper Novoalign and ChIP-Seq peak caller QuEST (Valouev et al. 2008). After performing peak calling relative to the input data and applying appropriate normalizations to the data, we found that Pol II binding to repeats showed an average 4.3-fold increase in *piwi* mutants (Figure 1F). In addition, 463 and 173 genes were increased and decreased for Pol II occupancy in the *piwi* mutant, respectively, most of which are involved in development (Figure 1G). All these analyses support Claim 6 by Huang *et al*. (2013) that Piwi influences Pol II distribution. However, differences in the sample quality of the ChIP-seq experiments in the wild-type and *piwi* mutant could also account for the observed enrichment. Additional experiments with replicate measurements and statistical differential analysis of peak enrichment will further confirm this bioinformatic result.

References

Chung D, Kuan PF, Li B, Sanalkumar R, Liang K, Bresnick EH, Dewey C, Keles S. Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. PLoS Comput Biol. 2011; 7:e1002111. [PubMed: 21779159]

Huang XA, Yin H, Sweeney S, Raha D, Snyder M, Lin H. A major epigenetic programming mechanism guided by piRNAs. Dev Cell. 2013; 24:502–516. [PubMed: 23434410]

Marinov GK, Wang J, Handler D, Wold BJ, Weng Z, Hannon GJ, Aravin A, Zamore PD, Toth KF, Brennecke J. Pitfalls of mapping high throughput sequencing data to repetitive sequences: Piwi's genomic targets still not identified. Dev Cell. 2015 *this issue.*

Sun G, Chung D, Liang K, Keles S. Statistical analysis of ChIP-seq data with MOSAiCS. Methods Mol Biol. 2013; 1038:193–212. [PubMed: 23872977]

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods. 2008; 5:829–834. [PubMed: 19160518]

Yin H, Sweeney S, Raha D, Snyder M, Lin H. A high-resolution whole-genome map of key chromatin modifications in the adult Drosophila melanogaster. PLoS Genet. 2011; 7:e1002380. [PubMed: 22194694]

Research Highlights

1. A direct response to the Marinov et al. (2015) Matters Arising is presented
2. New analysis indicates that Piwi genomic targets still remain to be identified
3. Three independent analyses support that Piwi influences Pol II distribution

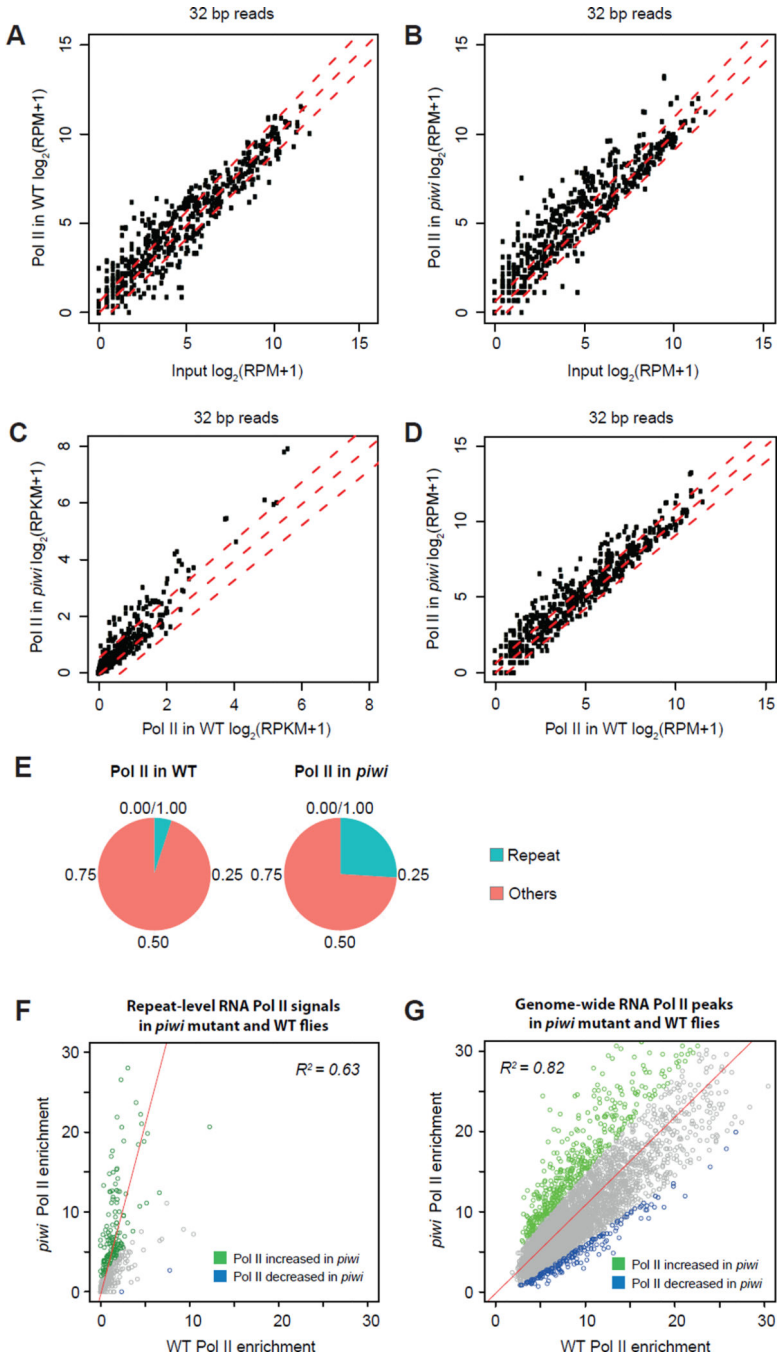

Figure 1. Impact of Piwi on RNA polymerase II distribution

(A-B) RPM scatter plots showing enrichment of Pol II signal in repeat sequences over input control in wildtype and *piwi* mutant flies, respectively. Each dot represents a repeat name, for which an RPM score is calculated for the quantification of abundance. Only reads unique to a specific repeat name were used in the calculation. X-axis represents the abundance of signal in the input control. Y-axis represents the abundance of Pol II signal in the wildtype sample. Red dashed lines represent fold change equal to 2, 1, and 0.5, respectively.

(C-D) RPKM and RPM scatter plots showing enrichment of Pol II signal in repeat sequences in *piwi* mutants. The calculation and presentation format are the same to (A-B). X-axis represents the abundance of Pol II signal in the wildtype sample. Y-axis represents the abundance of Pol II signal in the *piwi* sample. Red dashed lines represent fold change equal to 2, 1, and 0.5, respectively.

(E) Pie charts for annotations of Pol II binding peaks.

(F-G) RNA Pol II normalized signals across repeat classes and genomic peaks of Pol II, respectively, in *piwi* mutant and wildtype flies. The lines show a linear fit.